27.10.2025

Mikael Slotboom, Niilo Kurki, Tom Cordruwisch

Introduction to Data Science 2025

Group 3 - Project Report

1. Introduction:

The goal of this project was to help people live longer and healthier. We were interested in whether it would be feasible to make a tool that would predict a user's life expectancy as a concrete number and give suggestions on how to be healthier. Rather than population averages and individual risk/benefits studies which tend to be easy to brush off or disregard ("this doesn't apply to me"), our application would make it feel more personal and tailored, hopefully motivating the user more strongly to pursue positive life habits.
To do this, we would require a range of inputs from the user: personal information that the user would provide themselves, which would allow a machine learning model to run inference and return a prediction. Then, we could give personalised tips to the user on how to increase their life expectancy and what things they may want to prioritise.

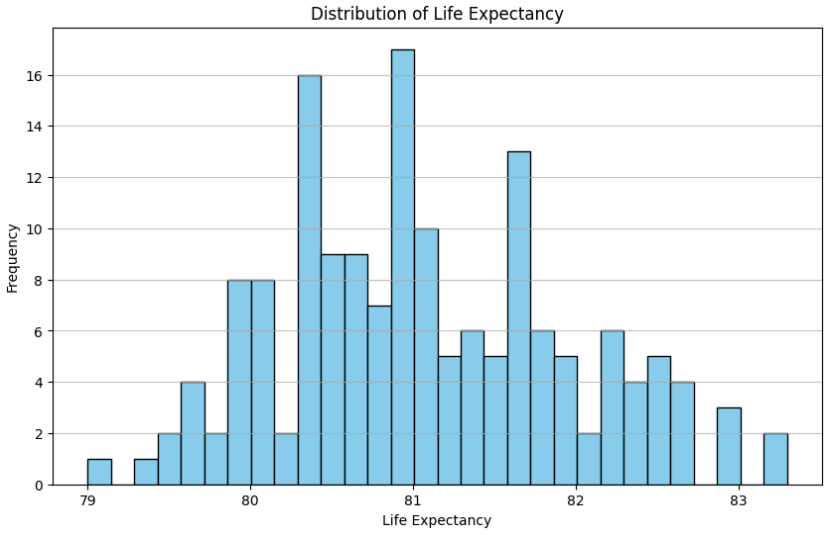
2. Data acquisition:

After deciding the topic of our project, we started searching for suitable datasets. Unfortunately, data acquisition proved challenging, one hindrance was that the access to many health-related datasets is restricted for privacy reasons. As a result, datasets with information about individuals (age, sex, habits) was not easily available to us. We adapted our strategy to instead look for data containing causes of death on a national level. This data was restricted, as it was only possible to download a specific number of rows at a time. To circumvent this, we developed a script to download data in chunks, and finally merged them into a single dataset. In the end, this data proved to not be very useful however, as it did not include information about the individuals backgrounds for the cause of death. This made it impossible to specifically predict someone's risk of death by cause. Furthermore, many of the causes of death were overly specific, and as a result of Finland being a country with a modest population size, most causes of death consisted of rather low values, i.e. zeroes or single-digit values. An additional concern was missing data, as the dataset relied on causes of death being properly recorded.
Moving forward with our search, we considered finding supplemental data about lifestyle factors that increase or decrease the risk of dying in specific ways. For instance, we tried researching how strongly smoking is linked to various types of disease, so we could subsequently connect it with our dataset containing the various causes of death. However, even after doing so, we were faced with the problem that we still lacked an actual way to translate this data into a concrete life expectancy and how much each habit affects would impact it. Furthermore, this strategy required a lot of extra effort, requiring detailed research about each specific life habit, which was surprisingly difficult. The extent to which lifestyle habits impact life expectancy is not trivial to quantify and depends on a lot of interlinking factors and correlations, even for more obvious causal links like smoking and alcohol consumption. On top of each data source having its own risk of bias and inaccuracies,

further risking the introduction of compounding errors, many of the factors we considered such as sleep, work, physical activity, mental health, obesity, access to health care, education, were hard to link to life expectancy directly. For these reasons, and since we hoped to target our main audience of people living in Finland, we ended up continuing our research. Finally we found data that we felt we could work with, sourced from the Finnish Institute for Health and Welfare (THL) and the national statistical institution Finland (Statistics Finland/Tilastokeskus). These institutions provide free access to diverse datasets about many different indicators concerning the Finnish population. In the end, we collected datasets for over 20 different indicators that we expected to have an impact on life, subdivided by year (2010 to 2023) and Finnish administrative region.

3. Data Exploration and pre-processing:

We combined the datasets sourced from THL and Tilastokeskus, containing the regional statistics, aiming to correlate them with the life expectancy of the respective regions. The statistics list different yearly health-related information, such as daily smokers, physical activity, alcohol sales and of course life expectancy. As the datasets from the two sources had different formats, we needed to make them consistent and merge them into a single dataframe, where each row represents a different year and region, while the columns correspond to the monitored indicator features.
The life expectancy-column constituted our labels (prediction targets), in other words, we were able to simply split the data into input features and life expectancy as output.



Histogram of life expectancy

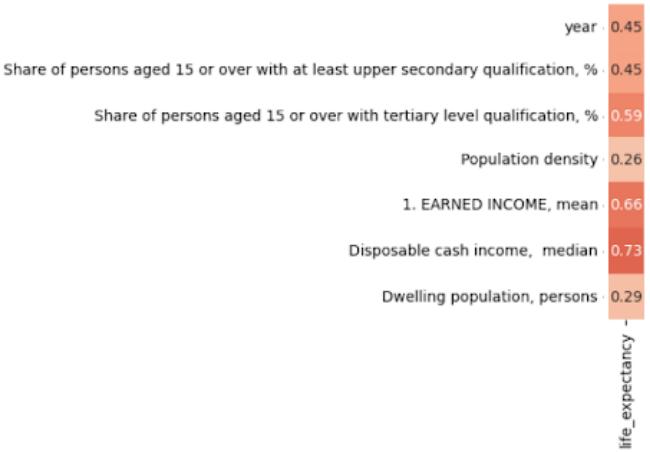
Looking at the distribution of life expectancy in our dataset, what stood out is the rather limited range: since the data is restricted to averages collected from populations of Finnish regions, it was clear that we likely would not be able to accurately predict life expectancy outside of the range represented in the data. Despite that, we decided that it was better to move ahead with the project, and try to tackle the resulting challenges rather than getting stuck on searching for datasets.

27.10.2025

Mikael Slotboom, Niilo Kurki, Tom Cordruwisch

To gain a bit more information we generated various plots, including the correlation heatmap which already gave us a considerable amount of insight for the correlation of life expectancy with the remaining features. As is typical with correlation plots, values range from -1.00 (perfect negative correlation) to 1.00 (perfect positive correlation). While some of our observations were expected, there were several surprises too and we noted the following things:

First, looking at some of the more positive correlations, the given year was moderately correlated with higher life expectancy, meaning that in the past ~15 years life expectancy has generally improved in Finland. Among categories that had the highest positive correlations were disposable cash income, earned income and share of the adult (15+) population with completed tertiary education. These correlations were moderately high to high and relatively unsurprising. It is generally expected that wealthier, more educated populations tend to have better access to high quality healthcare.



For negative correlations, evidently unhealthy indicators such as disability ratio and alcohol sales were unsurprisingly the most harmful factors. Some of the moderately negative correlations, like daily smokers and incidence of disability pensions we also expected. While daily smoking did not have a strong negative correlation, the data counted everyone who admitted to smoking daily in a survey, but did not further distinguish the type or amount of tobacco products consumed.

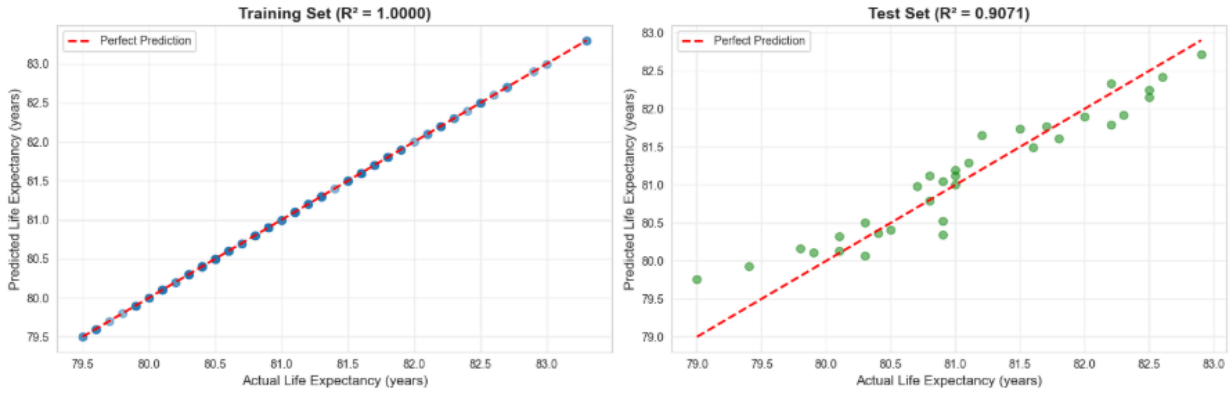
Factors that were less strongly correlated than we anticipated were physical activity, happiness, and obesity rate. We concluded that both obesity rate and happiness may not affect life expectancy immediately, since the time of death as a consequence of these indicators can lag behind considerably. Also, as the obesity rate has been considerably lower in the past, the full impact of the current rate is yet to substantially affect life expectancy in the present day. The same reasoning can be applied to other health indicators like physical activity and regular sports events. On the other hand, factors like mental health and “work until retired” (share of people who feel like they will not be able to work until retirement) have stronger correlations. This may be attributable to linked risk factors like suicide risk affecting average life expectancy much more significantly and immediately, compared to harmful behaviours with predominantly long-term consequences. This revealed a limitation in our data: factors with obvious positive/negative impacts may only notably affect life expectancy decades later. In addition, some aspects like happiness and physical activity can vary a lot over the span of peoples’ lives, making it difficult to create objective measures for describing long term physical and mental health of individuals.

Finally, we looked at some of the weakly and uncorrelated features, i.e. correlation coefficients close to 0. Surprising were the lack of impact of severe mental strain and share of rural/urban population on life expectancy. We would have expected notably higher life expectancy for people living in urban areas, due to better access to healthcare, and a clearly negative correlation for severe mental strain. Although this might have been a result of the data collection methodology, based on the indicator's description, the methodology appeared robust. We may assume that this observation could be related to short-term vs. long-term factors, or perhaps there are biases introduced by e.g. dissimilar age distribution or willingness to seek medical help.

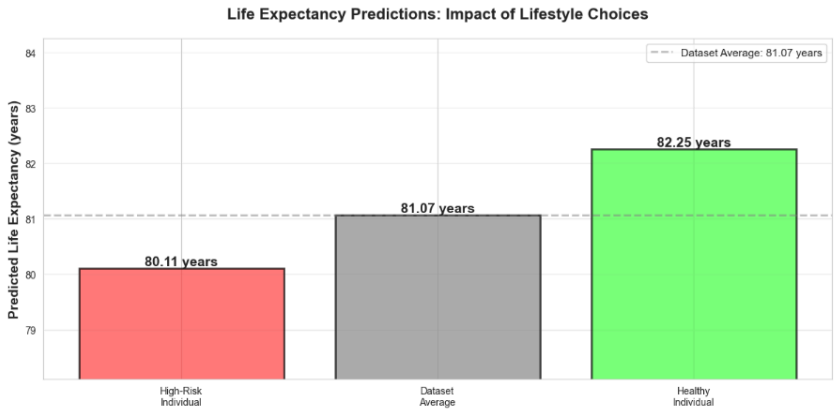
After this data exploration we felt that we had a good idea about the data we were working with and started preparing it for ML training. In order to do so, we performed some light pre-processing. We replaced missing values with the respective median which we felt was the most reasonable approach for data concerning a human population. We split the data according to an 80:20 train-test-split and applied feature normalisation using a standard scaler.

4. Implementation:

For our initial tests, we trained ML models using a selection of different algorithms, including RandomForest, a DNN regression model and XGBoost. We went ahead with XGBoost for our final implementation as it made the most robust predictions, achieving an R^2 score of around 90% on the test set.



After analysing the impact of features on the model predictions, we split them into positive (increase life expectancy) and negative (decrease life expectancy) features.
SHAP analysis helped us to further verify the kind of impact each feature has on life expectancy. Based on this, we identified key features to focus on, with the positive features being income, disposable cash, ratio of tertiary education and the negative features being disability, alcohol sales and daily smoking.
Based on our analysis, we determined the user inputs to base the application on, focusing on features with strong impact on predictive accuracy ("top features") and ones that are helpful for creating actionable suggestions for users for improving their life expectancy.
We tested the predictions with hypothetical user profiles following a high-risk lifestyle vs. a healthy lifestyle.



To create predictions based on user input however, the remaining challenge was who to translate population-wide values and ratios into metrics that could represent the life habits of an individual. To solve this problem, we applied feature conversions to obtain derived features: converting for instance a user's education level or smoking habits into numerical values aligning with the values present in the data. After the user input is converted and validated, it is used to generate a feature vector, overwriting the median placeholder values of the given features, and finally used for model inference. For giving personalised advice relevant for each prediction, we created a set of messages depending on whether the indicator is above or below the median value in the data.

While the resulting predictions are not perfect, we wanted to err on the side of caution with our approach to avoid giving outrageous predictions to users. Although the predictions are relatively conservative, likely underestimating the real-world variance, given the dataset this was the option that made the most sense for something as critical as personal health and longevity. Despite that, the personalised recommendations are meaningful and relevant to the user's background, and we feel that this better serves the main goal and usefulness of the project.

5. Takeaways:

The project has definitely taught us a lot, for example one aspect that we already anticipated to be very challenging, and subsequently experienced ourselves,was to find suitable datasets, collecting data and pre-processing it. Data analysis proved vital to inform our decisions on what features should be integrated into the frontend.
One major aspect we learned from the project is how very fundamental flaws (e.g. applying population data to individuals) can be easy to overlook in the early stages, but will inevitably materialise and make progress more difficult down the line. This meant that we had to come up with new strategies after a lot of work had already been committed. For that reason it makes sense why the software/AI development industry relies heavily on iterative development processes alongside well thought-out planning. Additionally, it has shown us how domain experience can be indispensable in projects for creating better and smarter solutions.